Un projet du CNRS en cours de réalisation : L'informatisation du Petit Larousse 1905 et d'une collection millésimée et séculaire

Helene Manuelian, Carine Timmermann

Un projet du CNRS en cours de réalisation : L'informatisation du *Petit Larousse 1905* et d'une collection millésimée et séculaire.

Hélène MANUÉLIAN[❖], Carine TIMMERMAN[✕]
[❖]Université de Cergy-Pontoise - UMR Métadif
[✕]CNRS - LACITO

Cet article a pour but de faire le point sur le projet d'informatisation du *Petit Larousse 1905*, et d'en expliciter les motivations. Nous revenons dans cet articles sur les raisons de l'informatisation des dictionnaires en général, et sur l'intérêt d'informatiser le Petit Larousse de 1905 en particulier. Nous exposons dans un deuxième temps les conséquences techniques des raisons d'informatiser un dictionnaire. Nous terminons par l'exposé du travail réalisé et en cours et les questions en suspens.

1. Les buts de l'informatisation

L'informatisation des dictionnaires est une pratique de plus en plus courante. Si on se réfère à J. Pruvost, les motivations sont au nombre de trois (Pruvost, 2000). Il s'agit de pouvoir disposer d'un support à la contenance quasiment illimité, et donc de pouvoir stocker un très grand nombre d'informations. La mise à jour d'un dictionnaire au format électronique est beaucoup plus faciles et moins onéreuse que la mise à jour d'un dictionnaire papier. Enfin, le type de requête permis par l'informatique permet d'approcher une utilisation analogique du dictionnaire rendue très difficile, même si elle existe, sur un support papier.

A ces trois intérêts, nous ajoutons l'idée qu'avec les nouvelles technologies, un dictionnaire informatisé peut aussi devenir un dictionnaire électronique au sens de M. Gross, c'est à dire un dictionnaire exploitable par une machine dans un but de traitement automatique des langues (Gross, 1975).

Concernant le Petit Larousse 1905, il est évident que les problèmes du stockage et de la mise à jour ne se posent pas, puisqu'il s'agit d'un support figé, lui même mis à jour en 1906 et toutes les années qui ont suivi. Cependant, les buts de consultation sur un support pratique, multimédia et analogique ainsi que l'utilisation pour le traitement automatique restent valables. Nous allons maintenant développer plus précisément les motivations pour l'informatisation d'un tel dictionnaire. Tout d'abord, nous souhaitons revenir sur les raisons qui rendent intéressante l'informatisation d'un dictionnaire comme le Petit Larousse 1905. Nous préciserons ensuite les buts exacts de notre informatisation.

1.1 Pourquoi le Petit Larousse 1905 ?

On peut en effet se demander quel est l'intérêt d'informatiser un dictionnaire comme le Petit Larousse de 1905. Il s'agit en effet d'un dictionnaire relativement petit, qui contient donc des informations relativement simplifiées. Par ailleurs, il s'agit d'un vieux dictionnaire, qui décrit un lexique déjà ancien, sans pour autant être totalement oublié. Nous voyons cependant plusieurs bonne raisons de faire ce travail.

1.1.1 Patrimoine scientifique et linguistique

La première raison est l'intérêt du Petit Larousse de 1905 en terme de patrimoine scientifique et linguistique. En effet, ce dictionnaire, au même titre que n'importe quel autre, présente une description du lexique de la langue française à une période donnée, et mérite donc d'être conservé en tant qu'objet faisant partie du patrimoine scientifique et linguistique de la France.

L'informatisation permettra de conserver sur un support durable des informations tant sur le lexique français que sur les méthodes lexicographiques employées à l'époque.

1.1.2 Texte libre de droits

Une autre bonne raison de choisir le petit Larousse de 1905 est liée au fait que ce texte est assez ancien pour être passé dans le domaine public. Il permet donc d'en autoriser la consultation libre, et gratuite, ce qui est un élément qui nous tient à cœur. En effet, selon nous, les bases de données linguistiques devraient être des ressources libres pour la communauté scientifique, de manière à ce qu'elles soient utilisables par tous à moindre frais. Ceci représente non seulement un intérêt pour ceux qui les utilisent, mais aussi pour ceux qui les créent. En effet, en permettant une consultation large de la ressource, nous nous donnons la possibilité d'obtenir des retours sur notre travail, et de l'améliorer, ainsi que d'améliorer les outils et les théories autour de l'informatisation des ressources linguistiques.

1.1.3 Dictionnaire illustré

Une autre caractéristique intéressante du Petit Larousse de 1905 réside dans la présence d'illustrations. On trouve déjà des dictionnaires au format électronique contenant des illustrations, mais à notre connaissance, les dictionnaires anciens informatisés n'en contiennent pas. Nous tenons pour notre part à conserver cette caractéristique dans le format électronique parce qu'elle a été l'une des raisons du succès de ce dictionnaire.

1.1.4 Premier d'une longue série

Le Petit Larousse est le premier d'une longue série. L'édition de 1905 est la plus ancienne, et ce dictionnaire sortira tous les ans, mis à jour pendant un siècle et probablement pendant encore longtemps. Notre idée est alors d'informatiser la collection complète, de façon à avoir une compilation informatisée du lexique de la langue française au XX^{ème} siècle. L'intérêt sera de pouvoir consulter rapidement les définitions à travers les décennies du XX^{ème} siècle, de pouvoir étudier l'évolution des définitions, ainsi que d'avoir une idée des dates d'apparition ou de disparition des mots dans le dictionnaire.

1.2 Un but à court terme : la consultation du dictionnaire

La consultation du texte contenu dans le petit Larousse est bien évidemment le premier but que nous fixons au travail d'informatisation. Nous souhaitons remplir les exigences minimales des utilisateurs, en permettant la navigation d'une définition à l'autre en utilisant des liens hypertexte, et de permettre la formulation de requêtes simples.

1.2.1 Possibilité de navigation

La navigation est déjà quelque chose que les utilisateurs font avec les dictionnaires papier en un volume, donc typiquement avec le Petit Larousse. Ce que nous appelons navigation est l'action de rechercher une définition suite à la lecture d'une autre, (rechercher par exemple la définition d'un mot contenu dans la définition que l'on a consultée au départ), ou encore, la lecture d'une définition se trouvant quelques mots avant ou après celle que l'on a choisi de consulter au départ.

Pour cela, nous souhaitons insérer des liens hypertexte sur les mots pleins composant les définitions (noms et verbes essentiellement), ce qui ne sera pas sans poser problème en cas d'utilisation de formes fléchies. L'utilisation d'un lemmatiseur sera probablement nécessaire pour automatiser le processus de balisage.

De façon à donner l'impression d'une lecture séquentielle comme dans un vrai dictionnaire, il est souhaitable de laisser apparaître à l'écran la liste des mots précédant et suivant le mot dont on consulte la définition.

1.2.2 Requêtes

Dans un deuxième temps, nous souhaitons mettre en place un système de requêtes simples (nous n'avons pas l'intention de faire le TLFI en abrégé), qui permettrait par exemple de retrouver les mots appartenant à un domaine précis, ou ayant tel ou tel mot dans leur définition (par exemple, quels sont les mots du PLI qui contiennent le terme animal ou mammifère ?)

1.3 Un but à long terme : un dictionnaire utilisable en TAL

Un des buts que nous fixons à l'informatisation du dictionnaire est de créer une ressource qui soit non seulement utilisable par des utilisateurs humains, mais aussi par une machine. En effet, les technologies actuelles et les architectures de bases de données récentes le permettent. La création d'une telle ressource est par ailleurs essentielle pour le français, parce qu'actuellement, il n'existe pas de ressource lexicale informatisée en libre accès.

S'il existe en effet de nombreuses ressources pour l'analyse et la génération automatique de textes, elles sont essentiellement liées à la phonétique, la morphologie et la syntaxe. Lorsqu'on parle de lexique en TAL, il s'agit bien souvent de lexiques-grammaires, c'est-à-dire de lexiques qui contiennent uniquement des informations morphosyntaxiques sur les termes. Par ailleurs, la plupart du temps, les ressources linguistiques pour le TAL ne sont pas libres. Ceci signifie que très souvent, les linguistes doivent les re-créer pour leurs recherches, et on ne compte aujourd'hui plus les grammaires ou les lexiques grammaires qui ne sont en général pas diffusés au sein de la communauté scientifique.

Nous souhaitons donc réaliser une base de donnée lexicale qui permettrait :

- L'extraction des relations sémantiques (hyponymie, hypéronymie, méronymie) de façon à permettre la génération ou l'analyse automatique de textes. Ce type de base de données est essentielle dans le cadre du traitement de la référence par des logiciels de traitement automatique (établissement des coréférences, génération de texte plus naturelle en permettant la nom répétition systématique des syntagmes nominaux et choix des déterminants en génération de texte)
- De créer des moteurs d'inférence pour faire du *question answering*, c'est à dire de créer des systèmes de dialogue qui permettent d'obtenir des renseignements précis, sur des sujets comme la réservation dans le domaine du tourisme, par exemple.

2. Les conséquences du double but

Les deux objectifs que nous poursuivons ont des conséquences aussi bien du point de vue des choix informatiques qui sont faits, que du point de vue du travail linguistique.

2.1 Du point de vue informatique : Nécessité de tenir compte des normes et standards

Notre volonté de diffusion et de partage des ressource a une conséquence directe en informatique : il est nécessaire que les formats que nous utilisons soient des formats standards (par opposition aux formats qu'on appelle propriétaires, qui ne sont utilisables qu'avec un seul système d'exploitation, ou une seule application, et qui nécessitent généralement l'achat de logiciels ou de machines onéreux.). Par ailleurs, l'intérêt d'utiliser des formats standards réside dans la volonté de conservation du patrimoine que nous affichons. En utilisant ces formats, nous avons la garantie qu'au fur et à mesure de l'évolution des technologies, nos fichiers seront lisibles par les nouveaux logiciels, ou en tous cas, possible à convertir pour être lus par les futurs logiciels. Ceci n'est absolument jamais garanti par les logiciels et les formats commerciaux. Nous allons donc maintenant détailler les normes et les standards que nous allons utiliser.

2.1.1 Standards de codage des fichiers

XML : Le premier élément important pour l'informatisation d'une ressource textuelle est de la convertir dans un format de fichier standard. Pour la rendre lisible et exploitable, il est nécessaire

que le texte soit balisé, c'est à dire qu'on y ait inséré des éléments correspondant à des indications sur le contenu du texte, que la machine puisse interpréter. Le format standard pour les balises est le format XML, qui n'est en fait pas un langage de balisage à proprement parler, mais un protocole de stockage et de gestion de l'information. Il fait partie d'une famille de technologies qui permettent d'effectuer le formatage de documents et l'extraction de données, et constitue une philosophie de gestion de l'information qui recherche un maximum d'utilité et de souplesse en organisant les données sous la forme la plus pure et la plus structurée. Il s'agit en réalité d' un ensemble de règles qui permettent le balisage du texte en autorisant une très grande liberté dans le choix des balises. Les balises vont avoir plusieurs fonctions illustrées dans le cadre ci – dessous :

Délimitation d'un fragment de texte :
```
<paragraph> texte du paragraphe </paragraph>
```
Indication du rôle d'un fragment de texte :
```
<salutation> bonjour ! </salutation>
```
Indication de la position d'un élément dans un texte
```
<title> Texte du titre </title> <paragraph> Texte du paragraphe </paragraph>
```
Imbrication des éléments les uns dans les autres
```
<chapter> <paragraph> Texte du premier paragraphe </paragraph>
<paragraph> Texte du deuxième paragraphe </paragraph>
<paragraph> Texte du troisième paragraphe </paragraph>
</chapter>
```
Indication des liens entre les fichiers
```
<graphique fileref = "sourire.pict"/>
```

Une balise se compose éventuellement de plusieurs éléments :

Le premier élément est un nom, qui est obligatoire, il va indiquer exactement la signification de la balise. Certaines balises vont pouvoir être augmentées de couples « attribut – valeur », c'est à dire d'indications donnant des informations supplémentaires sur le sens du fragment de texte balisé. Par exemple, dans la balise suivante, on trouvera les indications suivantes : le domaine auquel appartient le mot défini est l'agriculture. On entourera alors cette indication d'une balise <domain> pour signifier à la machine que le fragment de texte indique le domaine dans lequel le mot s'emploie. L'attribut « type » a pour valeur « agric », ce qui permet d'indiquer à la machine le sens du texte qui est compris entre les deux balises <domain>.

```
<domain type = agric> agriculture </domain>
```

XSL et HTML : HTML est le format standard d'un navigateur web. Il est donc nécessaire pour une consultation libre de notre dictionnaire en ligne de l'utiliser. XSL est un langage qui permet de transformer un fichier XML en fichier HTML, lisible sur le Web.

La TEI (Text Encoding Initiative) et les recommandations du comité de l'ISO TC37/SC4 :

Modèle de document : XML fournit ce qu'on appelle un *modèle de document,* qui est un ensemble de règles propres à un type de document (roman, dictionnaire, article de presse, etc.). Ces règles permettent de comparer le document produit à un document du même type et de dire s'il est conforme aux règles. On parlera alors de validation du document par le modèle. La plupart du temps, un modèle de document est ce qu'on appelle une DTD (Document Type Definition), mais on trouve aujourd'hui ce qu'on appelle des schémas XML. La DTD est un ensemble de règles qui indiquent quelles balises le document peut utiliser en fonction de sa nature. Elle fournit une description formelle de l'organisation de l'information au sein du document, la liste des attributs

possibles pour une balise et les valeurs possibles de ces attributs. On fait référence à la DTD utilisée au début du document pour que XML puisse valider le document. Les DTD peuvent être normalisées si on se réfère aux recommandations de la TEI.

La TEI est un projet international mis en place à la fin des années quatre-vingts dans le but de créer un environnement dans lequel les documents pourraient être encodés de façon à ce que leurs propriétés soient transcrites et que leur transcription puisse être échangée et survivre aux évolutions technologiques (Mueller, TEI, 2002).

Elle est en fait un format de représentation générique des ressources textuelles. Elle a permis de fournir une base commune pour la normalisation des documents, mais reste flexible. Les utilisateurs ont en effet la possibilité de choisir leur schéma de codage parmi les différents attributs qu'elle propose. Un document normalisé selon la TEI comporte cependant au moins deux éléments : l'en-tête et le texte qui constitue en lui même la ressource linguistique.

L'en-tête *TEI peut être vue comme une page de titre qui serait attachée à la version imprimée de la ressource* (Bonhomme, 2000). Elle contient quatre parties qui sont balisées ainsi :

<fileDesc> description du fichier </fileDesc>
<encodingDesc> description du codage </encodingDesc>
<profileDesc> profil textuel du document (classification du texte, thème, etc.) </profileDesc>
<revisionDesc> historique des changements </revisionDesc>

L'en-tête TEI permet donc de documenter les ressources comme l'auraient fait des bibliothécaires ou des documentalistes dans le fichier d'un centre de ressources. Son contenu a d'ailleurs été élaboré par des professionnels de la documentation (documentalistes, archivistes et bibliothécaires).

Concernant les dictionnaires, tout un chapitre de recommandation a été rédigé au sein du consortium TEI : il s'agit du chapitre 12 Print Dictionaries (Sperberg-McQueen, Burnard, 2004).

Parallèlement à la TEI qui est constituée d'experts du domaine affichant la volonté de normaliser les ressources textuelles, mais qui reste une initiative privée, on trouve un des sous comités de l'ISO (International Standard Organisation), le sous – comité 4 du comité technique 37 (désormais TC37/SC4), dont la fonction est à un niveau plus officiel et tout à fait aussi international, de valider les propositions de normalisation des ressources textuelles, et de publier des recommandations. Bien entendu, très vite, le TC37/SC4 a intégré les recommandations de la TEI dans ses normes.[1]

2.1.2 Standards de structuration des données

Les nombreux travaux d'informatisation de dictionnaires s'inscrivent dans la double perspective que nous décrivons au début de cet article : **(1) la modélisation des dictionnaire**s dans la lignée des travaux de Ide, Véronis et Lemaître qui ont donné naissance au chapitre 12 de la TEI cité précédemment, et **(2) la construction de bases de données** lexicales pour le TAL.

Actuellement, au travers du projet LMF, les comités de l'ISO ouvrent un projet de spécification de structure de bases de données lexicales et lexicographiques qui **unifie ces différents modèles** et dont l'architecture simple et extensible est représentée dans la figure 3.[2]

La norme LMF aura pour but de produire des formats standards pour tous les types de bases lexicales, dont les dictionnaires. Elle s'appuie sur les travaux menés dans la TEI et constituera une base de réflexion sur la façon de structurer les dictionnaires pour la prochaine version de la TEI (P5). Dans ce cadre, des propositions sont actuellement faites pour annoter et structurer les notes historiques des définitions du Trésor de la Langue Française Informatisé. Ces normes sont utilisées au niveau international, puisqu'elle sont réutilisées pour le GRIMM (dictionnaire d'allemand) et le Oxford English Dictionnary (pour l'anglais).

[1] http://www.tc37sc4.org

[2] Le projet de norme LMF est actuellement en cours d'élaboration dans le comité de normalisation ISO, au sein du sous comité dirigé par Laurent Romary (TC 37 / SC 4). Ce projet porte le numéro 24 613 et il est coordonné par Monte George et Gil Francopoulo.

Concernant l'informatisation du Petit Larousse depuis 1905, on peut envisager de spécifier et de faire des propositions aux comités de l'ISO et de la TEI, afin de définir une base commune d'encodage des éditions du Petit Larousse, et de définir des éléments spécifiques aux éditions, qui ont sans doute évolué en un siècle.

L'intérêt d'utiliser les normes existantes est alors double : il permet de construire la nouvelle ressource grâce à des formats d'échanges des données simples à utiliser, et il permet au reste de la communauté scientifique d'accéder aux données du Petit Larousse sans problèmes techniques. En effet, le coût énorme de l'informatisation (création et maintenance) de telles données nous poussent à dire qu'elles ne doivent pas être construites dans l'isolement, mais reliées à d'autres initiatives, de façon à bénéficier d'un enrichissement mutuel. Notre but sera donc de toujours viser une compatibilité totale du point de vue informatique, aussi bien en termes de logiciels que de systèmes d'exploitation

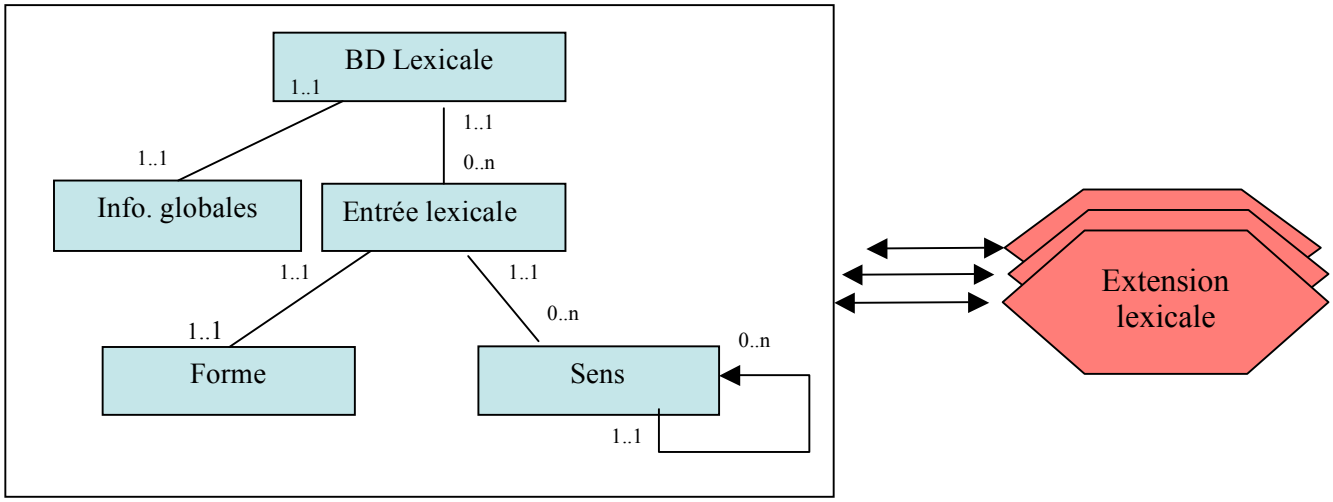


Fig. 3 : Architercture de LMF

2.2 Du point de vue linguistique : une analyse à deux niveaux

L'informatisation nécessite aussi une analyse linguistique du contenu du dictionnaire extrêmement précise. La première partie du travail concerne les lexicographes, puisqu'il s'agit de réussir à exprimer en termes métalexicographiques la structuration et la composition des entrées du dictionnaire. La seconde partie consiste à analyser les définitions au niveau sémantique, de manière à pouvoir les exploiter aussi bien en métalexicographie qu'en traitement automatique des langues. Ce sont ces deux étapes du travail que nous allons détailler maintenant.

2.2.1 Une analyse lexicographique du dictionnaire

Cette partie du travail a déjà été réalisée (cf. paragraphe 3.2.3) pour le Petit Larousse 1905. Il s'agit d'étudier les entrées une à une de manière à déterminer la liste des éléments qui les composent, et éventuellement les éléments de mise en forme qui leur correspondent.

L'énumération des différents composants des entrées (définitions, exemples, collocations, informations grammaticales, etc.), même s'ils ne sont pas présents dans toutes les entrées permet d'établir la liste de toutes les balises XML qui seront nécessaire au balisage du texte.

Ce travail descriptif est essentiel à la rédaction de la DTD, et à la compréhension du fonctionnement du dictionnaire, nécessaire à l'informatisation.

2.2.2 Une analyse sémantique approfondie des définitions

La seconde partie du travail consiste à analyser les définitions elles – mêmes, puisque ce sont elles qui vont constituer le fond de la base de données lexicale.

Pour le point de vue métalexicographique, il s'agira de fournir par exemple une analyse et un balisage des définitions pour différencier définitions logiques et définitions synonymiques.

Pour le traitement automatique des langues, il s'agira, après une analyse syntaxique et morphologique des définitions, d'établir des procédures de récupération des relations lexicales, pour permettre l'extraction d'ontologies par exemple. On pourra aussi fournir une analyse des verbes et des compléments qu'ils sous-catégorisent pour établir une base de donnée des cadres de restrictions de sélection.

3. Le travail réalisé

Les objectifs à court et à moyen terme étant définis, il est maintenant temps de faire le point sur l'état d'avancement du travail, depuis la numérisation du texte jusqu'à son « prébalisage ». L'informatisation, précisons-le, ne concerne pour le moment que la partie des noms communs.

3.1 La numérisation

Avant tout se pose le problème – que beaucoup sous-estiment mais qui reste d'importance – de la transposition du texte papier sur support numérique. Quelles sont les possibilités ? Si l'on regarde du côté de l'informatisation des dictionnaires anciens, comme le *Thresor de la langue françoyse* (1606) de Jean Nicot informatisé par Terence Russon Wooldridge[3], le *Dictionaire critique* (1787-1788) de Jean-François Féraud par Philippe Caron[4], l'*Oxford English Dictionary*[5] (1884-1928 pour la première édition)..., dans la majorité des cas, la saisie manuelle du texte est privilégiée : d'abord en raison de la fragilité de ces ouvrages et par conséquent de leur préciosité (il n'en reste bien souvent qu'un seul exemplaire) ; ensuite parce que cette méthode assure le respect au plus près de la reproduction des caractères spéciaux inhérents aux ouvrages anciens. L'inconvénient de cette méthode est évidemment les lourds moyens techniques et humains à mettre en œuvre, d'autant que souvent il est préférable de faire appel à des opérateurs de saisie étrangers, ne connaissant pas ou peu le français pour les préserver de toute tentative de surcorrection.

Toutefois, la technologie numérique permet maintenant de scanner ce type d'ouvrage sans « toucher » au support. Il suffit de voir pour cela l'immense travail de numérisation fait à la Bibliothèque nationale de France dans le cadre de l'archivage des textes (on saluera sur le sujet le site Gallica[6] qui permet entre autres de télécharger de nombreux ouvrages anciens, libres de droits).

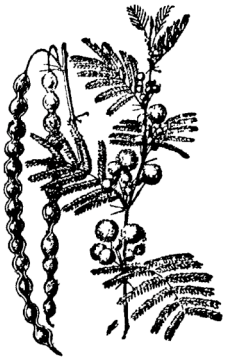
Du fait que le *Petit Larousse illustré* de 1905 n'est pas un ouvrage rare (les éditions Larousse ont ressorti à l'occasion du centenaire du *Petit Larousse illustré* le fac-similé de l'édition de 1905, malheureusement la numérisation s'est faite un an auparavant sur un ouvrage de l'époque), que son orthographe est moderne et que la qualité du papier est encore assez bonne (il n'est ni trop jaune, ni trop transparent), la numérisation est la méthode optimale de traitement, d'autant que notre équipe se réduit alors à trois personnes[7]. L'équipement faisant défaut (et le matériel de numérisation professionnel restant très coûteux), nous faisons appel à la société Azentis[8], spécialiste de la numérisation des œuvres anciennes, qui scannera la totalité de la partie des noms communs du *Petit Larousse illustré* en seulement trois jours (l'ouvrage est néanmoins massicoté pour être traité). Le dictionnaire est alors numérisé : chaque page donne lieu à un fichier tif, un fichier « image ». Pour que ceux-ci soient utilisables, ils doivent être convertis en fichiers textes grâce à un logiciel de reconnaissance de caractères optiques ou OCR (*Optical Character Recognition*). Chaque page est en même temps numérisée et convertie en texte ; on récupère alors les fichiers tif (Fig. 1) et les fichiers word (Fig. 2) sur cédérom.

---

[3] http://www.chass.utoronto.ca/~wulfric/nicot/nicot_tact.htm
[4] http://www.lib.uchicago.edu/efts/ARTFL/projects/dicos/FERAUD/
[5] http://dictionary.oed.com/
[6] http://gallica.bnf.fr/
7 L'équipe est à son début composée de Nicole Cholewka (Métadif), Anne-Marie Hetzel (Métadif) et Carine Timmerman (Lacito). Hélène Manuelian (Loria) nous rejoint en janvier 2005.
[8] www.azentis.com

ACABIT (*bi*) n. m. Qualité bonne ou mauvaise d'une chose : *poire d'un bon acabit. Fig.* et *fam.* Nature, caractère : *homme d'un bon acabit.*

ACACIA n. m. Arbre épineux de la famille des légumineuses, à fleurs odorantes disposées en grappes, et croissant dans les régions chaudes : *l'acacia de nos pays est le faux acacia ou robinier.*


Acacia d'Arabie.

ACADÉMICIEN (*si-in*) n. m. Autrefois, en Grèce, sectateur de Platon, dont l'école se tenait dans les jardins d'Académus. Aujourd'hui, membre d'une académie.

ACADÉMIE (*mî*) n. f. Société de gens de lettres, de savants ou d'artistes : *l'Académie française, l'Académie des sciences*, etc. V. ACADÉMIE (*part. hist.*). *L'Académie de médecine*, compagnie de médecins qui a son siège à Paris et qui ne fait pas partie de l'Institut. *L'Académie de musique* (à Paris), l'Opéra. Ecole de peinture, d'escrime, d'équitation. Division universitaire en France. — Il existe 17 académies (en comptant l'Algérie), dirigées chacune par un recteur assisté d'autant d'inspecteurs d'académie qu'il y a de départements dans sa circonscription. *Les 17 académies ont pour sièges : Aix, Alger, Besançon, Bordeaux, Caen, Chambéry, Clermont, Dijon, Grenoble, Lille, Lyon, Montpellier, Nancy, Paris, Poitiers, Rennes, Toulouse.*


Académicien.

ACADÉMIE (*mî*) n. f. Figure dessinée d'après un modèle nu.

ACADÉMIQUE adj. Propre à une académie : *fauteuil, séance académique. Style académique*, où l'art se fait trop sentir. *Pose académique*, prétentieuse.

ACADÉMIQUEMENT (*man*) adv. D'une manière académique : *traiter un sujet académiquement.*

ACADÉMISTE (*mis-te*) n. Personne qui tient une académie. Elève d'une académie.

ACAGNARDER (*gnar-dé*) v. a. Rendre fainéant. **S'acagnarder** v. pr. S'habituer à une vie oisive.

ACAJOU n. m. Arbre d'Amérique appartenant à des familles diverses, dont le bois est rougeâtre, très dur et susceptible d'acquérir un beau poli : *l'acajou, très employé en ébénisterie, prend une teinte rouge foncé en vieillissant.*


Acajou.

ACALÈPHES n. m. pl. *Zool.* Ordre de cœlentérés, comprenant les méduses, etc. S. un *acalèphe.*

ACANTHACÉES (*sé*) n. f. pl. Famille de plantes dicotylédones, dont l'*acanthe* est le type. S. une *acanthacée.*

ACANTHE n. f. (gr. *akantha*, épine). Plante épineuse du Midi, remarquable par ses feuilles très larges, élégamment découpées, recourbées et d'un beau vert. (Ses fleurs répandent une odeur forte et peu agréable.) Ornement d'architecture employé surtout sur les chapiteaux d'ordre corinthien, et qui imite cette plante : *feuille d'acanthe.*


Acanthe.

ACANTHIE (*ti*) n. f. Genre d'insectes dont l'espèce la plus connue est la punaise des lits.

ACANTHOPTÉRYGIENS (*ji-in*) n. m. pl. Famille de poissons ayant la nageoire dorsale épineuse, comme la *perche*, le *maquereau*, etc. S. un *acanthoptérygien.*

ACARIÂTRE adj. D'une humeur fâcheuse, aigre, criarde. ANT. **Doux, sociable.**

ACARIÂTRETÉ n. f. Humeur acariâtre.

ACARIDE ou **ACARIEN** (*ri-in*) n. m. Genre d'arachnides non articulés et souvent parasites.

ACARPE adj. Se dit d'une plante privée de fruit.

ACARUS (*russ*) n. m. Syn. ACARIDE.

ACATALECTIQUE (*lèk*) adj. Se dit, en métrique ancienne, d'un vers auquel ne manque aucune syllabe.

ACATALEPSIE (*lèp-sî*) n. f. Dans la philosophie grecque, impossibilité d'arriver à la certitude.

ACATÈNE adj. et n. f. (de *a* priv. et lat. *catena*, chaîne). Sans chaîne : *bicyclette acatène. Une acatène.*

ACATHOLIQUE adj. Se dit des chrétiens qui repoussent l'autorité du pape et de l'Eglise romaine.

ACAULE (*kô-le*) adj. (du gr. *a* priv., et *kaulos*, tige). Se dit d'une plante qui n'a pas de tige apparente, comme le *pissenlit*, le *plantain*, etc.

ACCABLANT (*a-ka-blan*), **E** adj. Difficile à supporter, qui accable : *poids accablant ; chaleur accablante. Fig.* : *chagrin accablant.*

ACCABLEMENT (*a-ka-ble-man*) n. m. Prostration physique ou morale. Extrême abattement.

ACCABLER (*a-ka-blé*) v. a. Faire succomber sous le poids. *Fig.* Surcharger : *accabler de travail.* Combler : *accabler d'honneurs.*

ACCALMIE (*a-kal-mî*) ou plus rarem. **ACCALMÉE** (*a-kal*) n. f. *Mar.* Calme momentané du vent et de la mer. *Fig.* Temps de repos momentané, après une période d'activité ou d'agitation.

ACCAPAREMENT (*a-ka, man*) n. m. Action d'accaparer, de prendre tout pour soi. — L'accaparement est puni comme un crime commercial ; il consiste à retirer de la circulation une forte quantité de denrées ou marchandises de même espèce, afin d'en avoir le monopole et de pouvoir, en écartant toute concurrence, déterminer la hausse ou la baisse des prix.

ACCAPARER (*a-ka-pa-ré*) v. a. (préf. *ad*, et lat. *capere*, prendre). Amasser une denrée quelconque en grande quantité pour en produire la rareté et la revendre fort cher. *Fig.* Prendre pour soi au détriment des autres : *l'empereur Auguste accaparatous les pouvoirs. Accaparer quelqu'un*, l'attirer sans cesse près de soi.

ACCAPAREUR, EUSE (*a-ka, eu-ze*) n. Celui, celle qui accapare : *accapareur de denrées, de faveurs.*

ACCASTILLAGE (*a-kas-ti, ll* mll.) n. m. Partie du vaisseau qui est hors de l'eau.

ACCASTILLER (*a-kas-ti, ll* mll., *é*) v. a. Garnir un navire de son accastillage.

ACCÉDER (*ak-sé-dé*) v. n. (lat. *accedere*, s'approcher. — Se conj. comme *accélérer.*) Avoir accès dans un lieu, arriver, parvenir. Adhérer, consentir, acquiescer. ANT. **Rejeter, refuser.**

ACCÉLÉRATEUR, TRICE (*ak-sé*) adj. Qui accélère, précipite : *la force accélératrice est directement proportionnelle à la masse mise en mouvement.*

ACCÉLÉRATION (*ak-sé, si-on*) n. f. Augmentation de vitesse qu'acquiert un corps en mouvement : *accélération du pouls.* Prompte exécution : *accélération des travaux.* ANT. **Ralentissement.**

ACCÉLÉRER (*ak-sé-lé-ré*) v. a. (Prend un *è* ouvert devant une syllabe muette : *j'accélère ;* excepté au fut. et au cond., où il conserve l'*é* fermé : *j'accélérerai, nous accélérerions.*) Hâter, presser, activer : *accélérer le pas.* ANT. **Ralentir, modérer.**

ACCENSE (*ak-san-se*) n. m. Chez les Romains, citoyen qui, d'après la constitution de Servius Tullius, n'atteignait pas les cens de la dernière classe. Appariteur attaché à la personne des magistrats.

ACCENSER (*ak-san-sé*) v. a. *Dr. anc.* Donner, prendre à cens une propriété.

ACCENT (*ak-san*) n. m. (préf. *ad*, et lat. *cantus*, chant). Elévation ou abaissement de la voix sur certaines syllabes : *accent oratoire.* Prononciation particulière : *accent gascon.* Expression de la voix : *accent plaintif.* Signe qui se met sur une voyelle : *il y a trois accents en français : l'accent aigu* (´), qui se met sur la plupart des *é* fermés : *bonté, café ; l'accent grave* (`), qui se met sur les *è* ouverts : *père, mère*, sur *où* (adv.), *à* (prép.), *holà, déjà*, etc. ; et *l'accent circonflexe* (^), qui se met sur les voyelles longues : *pâte, fête, gîte, côte, flûte. Fig.* Intensité de touche dans la peinture.

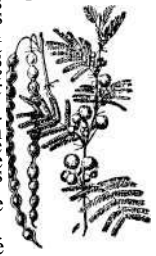
ACCENTEUR (*ak-san*) n. m. Genre d'oiseaux passereaux, qui vivent surtout dans les montagnes.

ACCENTUABLE (*ak-san*) adj. Qui peut être accentué : *mot accentuable ; syllabe accentuable.*

ACCENTUATION (*ak-san, si-on*) n. f. Manière d'accentuer, de prononcer, en parlant ou en écrivant : *accentuation vicieuse, faute d'accentuation.*

ACABIT *(bi)* n. m. Qualité bonne ou mauvaise d'une chose : *poire d'un bon acabit. Fig.* et *fam.* Nature, caractère : *homme d'un bon acabit.*

ACACIA n. m. Arbre épineux de la famille des légumineuses, à fleurs odorantes disposées en grappes, et croissant dans les régions chaudes : *l'acacia de nos pays est le faux acacia ou robinier.*



Acacia d'Amérique.

ACADÉMICIEN *(si-in)* n. m. Autrefois, en Grèce, sectateur de Platon, dont l'école se tenait dans les jardins d'Aca-démus. Aujourd'hui, membre d'une académie.

ACADÉMIE *(mi)* n. f. Société de gens de lettres, de savants, ou d'artistes : *l'Académie française, l'Académie des sciences,* etc. V. ACADÉMIE. *(part-hist.). L'Académie de médecine,* compagnie de médecins qui a

^.ll .IPf so

n siège à Paris et qui ne fait pas partie de l'Institut. *V Acac*

Paris), l'Opéra. Ecole de peinture, d'escrime, d'équita-tion. Division universitaire en France.— ^IS. Il existe 17 académies. (en comptant l'Algérie), dirigées chacune par un recteur assisté d'autant d'inspecteurs d'académie qu'il y a de départements dans sa circonscription. *Les 17 académies ont pour sièges : Aix, Alger, Besançon, Bordeaux, Caen, Chambery, Clermont, Dijon, Grenoble, Lille, Lyon, Montpellier, Nancy, Paris, Poitiers, Rennes, Toulouse.*

Académicien.

ACADÉMIE *(mi)* n. f. Figure dessinée d'après un modèle nu.

ACADÉMIQUE adj. Propre à une académie : *fauteuil, séance académique. Style académique,* où l'art se fait trop sentir. *Pose académique,* prétentieuse.

ACADÉMIQUEMENT *(man)* adv. D'une manière académique : *traiter un sujet académiquement.*

ACADÉMISTE *(mis-te)* n. Personne qui tient une académie. Elève d'une académie.

ACAHNAitDER *(gnar-dé)* v. a. Rendre fainéant. **S'apoRnarder** v. pr. S'habituer à une vie oisive.

ACAJOU n. m. Arbre d'Amérique appartenant à des familles diverses, dont le bois est rougeâtre, très dur et susceptible d'acquérir un beau poli : *l'acajou, très employé en ébénisterie, prend une teinte rouge foncé en vieillissant.*

Acajou.

ACALEPHES n. m. pl. *Zool.* Ordre de coelentérés, comprenant les méduses, etc. S. *méduse.*

ACANTHACÉES *(se)* n. f. pl. Famille de plantes dicotylédones, dont *l'acanthe* est le type. S. une *acanthacée.*

ACANTHE n. f. (gr. *akantha,* épine). Plante épineuse du Midi, remarquable par ses feuilles très larges, élégamment découpées, recourbées et d'un beau vert. (Ses fleurs répandent une odeur forte et peu agréable.) Ornement d'architecture employé surtout sur les chapiteaux d'ordre corinthien, et qui imite cette plante : *feuille d'acanthe.*

ACANTHIE *(ti)* n. f. Genre d'insectes dont l'espèce la plus connue est la punaise des lits.

ACANTHOPTERYGIENS *(ji-in)* n. m. pl. Famille de poissons ayant la nageoire dorsale épineuse, comme la *perche,* le *maquereau,* etc. S. un *acanthoptérygien.*

ACARIÂTRE adj. D'une humeur fâcheuse, aigre, criarde. ANT. DOUX, sociable.

ACARIÂTRETÉ n. f. Humeur acariâtre.

ACAHIDE ou **ACARIEN** *(ri-in)* n. m. Genre d'arachnides non articulés et souvent parasites.

ACARPE adj. Se dit d'une plante privée de fruit.

ACARUS *(russ)* n. m. Syn. ACARIDE.

ACATALECTIQUE *(lèk)* adj. Se dit, en métrique ancienne, d'un vers auquel ne manque aucune syllabe.

ACATALEPSIE *(lep-si)* n. f. Dans la philosophie grecque, impossibilité d'arriver à la certitude.

ACATÈNE adj. et n. f. (de *a* priv. et lat. *catena,* chaîne). Sans chaîne : *bicyclette acatène. L'acatène.*

ACATHOLIQUE adj. Se dit des chrétiens qui repoussent l'autorité du pape et de l'Eglise romaine.

ACAULE *(ko-le)* adj. (du gr. *a* priv., et *kaulos,* tige). Se dit d'une plante qui n'a pas de tige apparente, comme le *pissenlit,* le *plantain,* etc.

ACCABLANT *(a-ka-blan),* E adj. Difficile à supporter, qui accable : *poids accablant ; chaleur accablante. Fig. : chagrin accablant.*

ACCABLEMENT *(a-ka-ble-man)* n. m. Prostration physique ou morale. Extrême abattement.

ACCABLER *(a-ka-blé),* v. a. Faire succomber sous le poids. *Fig.* Surcharger : *accabler de travail.* Combler : *accabler d'honneurs.*

ACCALMIE *(a-kal-mi)* ou plus rarem. **ACCAL-MEE** *(a-kal)* n. f. *Bar.* Calme momentané du vent et de la mer. *Fig.* Temps de repos momentané, après une période d'activité ou d'agitation.

ACCAPAREMENT *(a-ka-man)* n. m. Action d'accaparer, de prendre tout pour soi. — L'accaparement est puni comme un crime commercial; il consiste à retirer de la circulation une forte quantité de denrées ou marchandises de même espèce, afin d'en avoir le monopole et de pouvoir, en écartant toute concurrence, déterminer la hausse ou la baisse des prix.

ACCAPARER *(a-ka-pa-ré)* v. a. (préf. *ad,* et lat. *ca-pere,* prendre). Amasser une denrée quelconque en grande quantité pour en produire la rareté et la revendre fort cher. *Fig.* Prendre pour soi au détriment des autres : *l'empereur Auguste accaparatous les pouvoirs. Accaparer quelqu'un,* l'attirer sans cesse près de soi.

ACCAPAREUR, EUHE *(a-ka, eu-ze)* n. Celui, celle qui accapare : *accapareur de denrées, de faveurs.*

ACCASTILLAGE *(a-kas-ti, II* mil.) n. m. Partie du vaisseau qui est hors de l'eau.

ACCASTILLER *[a-kas-ti, II* mil., *éjv.* a. Garnir un navire de son accastillage.

ACCÉDER *(ak-sé-dé),* v. n. (lat. *accedere,* s'approcher. — Se conj. comme *accélérer.)* Avoir accès dans un lieu, arriver, parvenir. Adhérer, consentir, acquiescer. ANT. Rejeter, refuser.

ACCÉLÉRATEUR, TRICE *(ak-sé)* adj. Qui accélère, précipite : *la force accélératrice est directement proportionnelle à la masse mise en mouvement.*

ACCÉLÉRATION *(ak-sé, si-on)* n. f. Augmentation de vitesse qu'acquiert un corps en mouvement : *accélération du pouls.* Prompte exécution : *accélération des travaux.* ANT. Ralentissement.

ACCÉLÉRER *(ak-sé-lé-ré)* v. a. (Prend un *è* ouvert devant une syllabe muette : *j'accélère;* excepté au fut. et au cond., où il conserve l'*é* fermé : *j'accélérerai, nous accélérerions.)* Hâter, presser, activer : *accélérer le pas.* ANT. **Ralentir,** modérer.

ACCENSE *(ak-san-se).* n. m. Chez les Romains, citoyen qui, d'après la constitution de Servius Tullius, n'atteignait pas les cens de la dernière classe. Appariteur attaché à la personne des magistrats.

ACCENSER *(ak-san-sé)* v. a. *Dr. anc.* Donner, prendre à cens une propriété.

ACCENT *(ak-san)* n. m. (préf. *ad,* et lat. *cantus,* chant). Elévation ou abaissement de la voix sur certaines syllabes : *accent oratoire.* Prononciation particulière : *accent gascon.* Expression de la voix : *accent plaintif.* Signe qui se met sur une voyelle :

il y a trois accents en français : l'accent aigu (´), qui se met sur la plupart des *é* fermés : *bonté, café; l'accent grave* (`), qui se met sur les *e* ouverts : *père, mère,* sur *où* (adv.), *à* (prép.), *holà, déjà,* etc. ; ex. *l'accent circonflexe* (^), qui se met sur les voyelles longues : *hâle, fête, gîte, côte, flûte. Fig.* Intensité de touché dans la peinture.

ACCENTEUR *(ak-san)* n. m. Genre d'oiseaux passereaux, qui vivent surtout dans les montagnes.

ACCENTUABLE *(ak-san)* adj. Qui peut être accentué : *mot accentuable; syllabe accentuable.*

ACCENTUATION *(ak-san, si-on)* n. f. Manière d'accentuer, de prononcer, en parlant ou en écrivant : *accentuation vicieuse, faute d'accentuation*

Des deux figures qui précèdent, il faut noter que :

- le texte, une fois le logiciel d'OCR passé, est de qualité variable : certains mots sont parfaitement restitués, d'autres mal et d'autres encore sont manquants. La phase de nettoyage aura pour objectif de tout restituer à l'identique du texte papier ;
- la mise en page est conservée telle quelle, en deux colonnes avec filet séparateur ;
- les illustrations sont numérisées en même temps que le texte (en plus ou moins bonne qualité selon les cas, voir plus loin) ; un second passage paramétré spécialement pour les illustrations ne sera donc pas nécessaire.

La lecture du texte word sur la figure 2 est difficile notamment en raison de l'interlignage qui y est très serré ; les informations de mise en page, même si elles sont conservées dans les grandes lignes, relèvent ici plus du handicap. Dès que nous essayons d'y appliquer un quelconque réglage, simplement pour l'interlignage par exemple, le texte chasse aléatoirement, les données formant un complexe méli-mélo. Les illustrations, quant à elles, sont plus ou moins indépendantes du texte (nous en reparlerons plus loin quand nous traiterons le cas particulier des illustrations), il n'est donc pas possible de travailler sur cette version word, bien trop « polluée » par nombre de signes inutiles. Pour le nettoyage, les fichiers word seront convertis en fichiers « texte seulement », débarrassés de toute information de mise en page inutile à l'informatisation.

3.2 Le nettoyage du texte numérisé

3.2.1 Relecture et correction du texte

Si l'on regarde de plus près les deux figures ci-dessus, on comprend mieux le travail du logiciel d'OCR : celui-ci, lorsqu'il balaie le texte, compare chaque caractère avec sa liste de caractères connus qu'il retranscrit en fonction. C'est pourquoi il arrive que souvent un "r" se confonde avec un "n", un "li" avec un "t" (le point du "i" portant à confusion), etc. Malheureusement toutes ces fautes sont loin d'être régulières et ne peuvent ainsi faire l'objet d'un traitement automatique ; en corriger certaines reviendrait à en ajouter d'autres. Toutefois, lors de la relecture manuelle, Nicole Cholewka relève certaines fautes récurrentes dans des groupes de trois ou quatre lettres qui sont traitables automatiquement, sans risque d'ajout involontaire de fautes. Par exemple :

– "II mil"	*à transformer en*	"ll mll"
– "Fiq."	" "	"Fig."
– "PI."	" "	"Pl."
– "{"	" "	"("
– "y."	" "	"v."

Cette liste est certes minime par rapport à toutes les corrections nécessaires, mais elle a le mérite d'être fiable.

À la relecture, on constate que le texte, après OCR, est bon à 50 %. Après un an de nettoyage avec deux personnes à mi-temps, on comptabilise près de la moitié du texte relue (environ 700 pages de la partie des noms communs). Pour le moment, le nombre de relectures est fixé à trois : la première, actuelle, de nettoyage en profondeur par comparaison avec la version papier, la deuxième au cours du balisage et la troisième, ultime relecture de vérification qui concernera à la fois le texte et les balises. L'objectif au terme de ces trois relectures est de ne pas dépasser un quota d'une à deux fautes par page, sans quoi seront programmées de nouvelles phases de

nettoyage.[9] L'informatisation n'a aucune valeur tant que le texte informatisé n'est pas identique au plus près au texte papier. Une troisième personne s'ajoutera à l'équipe de relecture actuelle ; cette première phase devra être terminée fin 2005-début 2006.

3.2.2 Le cas des illustrations

L'intérêt du *Petit Larousse illustré*, comme son nom l'indique, réside dans le florilège des illustrations. La numérisation permet de récupérer d'un seul coup texte et illustrations, mais comme pour le texte, celles-ci sont plus ou moins parfaites. Pour s'en rendre compte, elles sont examinées une par une et isolées dans un fichier word nommé par la vedette correspondante. Environ trois cents d'entre elles sont à rescannées, soit parce qu'elles sont de mauvaise qualité, soit parce qu'elles comportent du texte qui, lui, est de mauvaise qualité, soit parce qu'elles n'ont pas été récupérées par le scanner.

Pour les illustrations de qualité correcte, le traitement sera le suivant :

– enregistrement sous un format d'image (jpeg) ;

– traitement sous un logiciel de retouche d'images (éclaircissement, travail des contrastes, etc.) afin d'obtenir une qualité optimale ;

– pour les illustrations avec légendes intégrées, traitement sous un logiciel de retouche d'images en vue de récupérer le texte et mise en place de pointeurs (les légendes seront ainsi interrogeables).

La phase de balisage prévoit d'insérer un pointeur à chaque vedette concernée par une illustration.

3.3 Avant le balisage XML, l'analyse lexicographique du texte

Au cours de son séminaire de DEA en 2004, Jean Pruvost, en tant que spécialiste laroussien, procède à une analyse minutieuse du *Petit Larousse illustré* de 1905, en vue de la mise en place d'un prébalisage du texte, sorte d'analyse lexicographique de la macrostructure du dictionnaire. Se joint à l'équipe Marine Lesprit, titulaire d'un DESS de lexicographie à Lille, pendant deux mois.

De l'analyse de Jean Pruvost, se dégage tout un tableau de marqueurs lexicographiques qui concernent tant la macro- que la microstructure. Les articles y sont finement décrits, de l'adresse monoforme à l'adresse polyforme, de l'exemple forgé simple à l'exemple forgé connoté, de l'illustration simple à l'illustration complexe ; cette analyse annonce l'objectif de l'informatisation : un balisage très fin qui permettra de nombreuses requêtes utiles au lexicographe.

3.3.1 Éléments d'analyse

Dans le *Petit Larousse illustré* de 1905, l'article est composé des six parties suivantes, dans l'ordre :

– l'entrée ;
– la transcription phonétique ;
– les indications grammaticales ;
– l'étymologie ;

---

[9] Pour avoir un ordre d'idée, on estime qu'« il y a relecture absolue lorsque se sont succédé 17 relecteurs différents » (J. Pruvost, 2000 : 118).

– le domaine de spécialité ;

– le groupe définitionnel (définition(s), exemple(s), poverbe(s), commentaire(s) encyclopédique(s), etc.).

L'analyse lexicographique du *Petit Larousse illustré* est rapidement présentée ici, dans ses grands traits, l'objectif étant surtout de bien expliquer par la suite quels sont les fondements du « prébalisage » et d'en justifier ses balises.

En gras et en majuscule, l'adresse ouvre l'article. Rien d'original là-dedans, si ce n'est le fait qu'elle connaît des formes allant de la plus simple, monoforme, à la plus complexe, l'adresse polyforme (exemples : « MUTIN, E », « FIORD ou FJORD »).

L'adresse présente différentes formes par déclinaisons du genre ou du nombre. La déclinaison du genre, dans sa forme simple, se résume à l'ajout de la marque minimale du féminin ou à l'ajout d'une marque morphologique du féminin spécifique (« FANFARON, ONNE ») ; sa forme complexe en est l'interruption phonétique (« FARAUD (*rô*), E »). L'adresse polyforme par déclinaison du nombre peut être là aussi simple ou composée. Pour le cas simple, soit il n'y a pas de déclinaison ou seulement une déclinaison double (« JOURNALIER (*li-é*), ERE »), soit, pour le cas complexe, l'adresse présente un mixte du genre et du nombre (« FATAL, E, ALS »).

Les variantes de l'entrée sont aussi de plusieurs ordres : qu'elles soient orthographique, morphologique ou compositionnelle, l'adresse ne manque pas de les lister. La variante orthographique est simple comme pour l'entrée « FAROUCH ou FAROUCHE », elle se situe en finale du mot ou en son milieu, ou même en son initiale (« HOPLOMACHIE ou OPLOMACHIE »). Elle connaît une forme complexe que l'on distingue dans les entrées telles que « FAQUIR (*kir*) ou FAKIR » ou « FASEYER (*zé-i-é*), FASIER (*zi-é*) ou FASILLER ».

La variante morphologique se rencontre aux entrées comme « DECRUMENT (*man*), DECRUSAGE, DECREUSAGE (*za-je*) ou DECRUSEMENT », « MYCOLOGIE et MYCETOLOGIE » ; sans oublier l'adresse à variante compositionnelle que l'on retrouve avec « JACK ou UNION-JACK ».

Enfin, d'autres formes d'adresse sont celles avec interruption par le genre (« JABLOIR n. m., JABLOIRE ou JABLIERE ») ou, plus rares, avec interruption étymologique (« GLORIA PATRI (mots lat. signif. *gloire au Père*, ou par abrév. GLORIA) n. m. »), ou encore des formes hautement complexes comme « CZAR (*kzar*) n. m. V. TSAR. – CZAREWITCH (*kza*) n. m. V. CESAREVITCH. – CZARIEN, ENNE (*kza-ri-in, è-ne*) adj. V. TSARIEN. – CZARINE (*kza*) n. f. V. TSARINE. »

L'analyse s'applique de même à l'examen des exemples, puis des illustrations. Brièvement sur les exemples, on remarque qu'ils sont souvent derrière les deux points en italique et qu'ils sont séparés entre eux par un point-virgule. On les rencontre aussi derrière un point, commençant par une majuscule. Sur le fond, les exemples sont parfois des proverbes ou dictons à vocation moralisatrice (sous ABREUVOIR, « les abreuvoirs doivent toujours être propres »). L'exemple est rarement neutre dans le *Petit Larousse illustré* : nombre d'entre eux sont ironiques, humoristiques ou même cyniques, l'auteur y faisant aussi part de remarques personnelles.

Tantôt conservateur, tantôt novateur, l'exemple sert à illustrer la norme linguistique de l'époque, éventuellement à la nuancer, l'ouvrir ou la diversifier. Il applique la définition (parfois en termes grammaticaux), la complète de manière linguistique, encyclopédique, etc. Les exemples sont omniprésents dans le dictionnaire, le mot doit tout le temps être présenté en situation, souvent d'ailleurs il est actualisé dans un bel emploi à vocation didactique : dans ce sens, beaucoup d'exemples sont à caractères culturel, historique, géographique, scientifique, littéraire… Le lexicographe est avant tout au service de la langue, preuve en est la floraison d'exemples panchroniques (neutres) qui ne sont là que pour illustrer un usage.

Les illustrations sont très nombreuses dans le *Petit Larousse illustré*, elles en seront d'ailleurs la marque de fabrique. Elles prennent place dans l'article lui-même, à côté de l'article, ou font l'objet de planches. Illustrations monofigurale ou polyfigurale, partielle ou totale, paradigmatique ou syntagmatique, scalaire ou ascalaire, anaphorique : toute la typologie des illustrations se retrouve dans le *Petit Larousse illustré* de 1905. Pour certaines, plus que de simples dessins, elles jouent un rôle dans l'explication du sens en apportant des informations terminologiques par exemple – certains mots ne se retrouvant pas ailleurs dans le dictionnaire –, ou en présentant la même chose sous des angles de vue différents (« PORC », Fig. 3), en mettant en place une typologie (différentes sortes de croix, Fig. 4), ou encore en illustrant les différentes étapes d'un processus (« BATRACIENS », Fig. 5).

Fig. 3 – Illustrations sous PORC

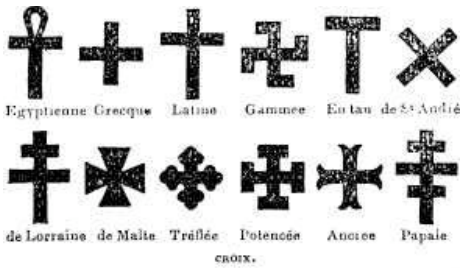


Fig. 4 – Illustrations sous CROIX

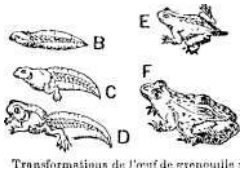


Fig. 5 – Illustrations sous BATRACIENS (« Transformations de l'œuf de grenouille »)

Ce sont ces éléments d'analyse (ici résumés) qui permettent à l'équipe l'élaboration d'un « prébalisage » lexicographique.

3.3.2 Le prébalisage

Le prébalisage est lancé avec dans l'idée de faciliter par la suite le balisage XML, et de permettre ainsi une étape supplémentaire de relecture mais en y ajoutant des éléments d'analyse qui facilitent la bonne connaissance du dictionnaire.

Voici le tableau simplifié des prébalises :

entrée	**<ENT></ENT>**
prononciation	**<PR></PR>**
catégorie grammaticale	**<CG></CG>**
conjugaison	**<CJ></CJ>**
étymologie	**<ETY></ETY>**
renvoi	**<RV></RV>**
renvoi > antonyme	**<ANT></ANT>**
renvoi > synonyme	**<SYN></SYN>**
renvoi > illustration	**<RIL></RIL>**
domaines de spécialité (à recenser au fur et à mesure du nettoyage) :	**3 premières lettres, ex. : <MUS></MUS> <ECC></ECC>...**
niveau de langue	**<NIV></NIV>**
fréquence	**<FQ></FQ>**
définitions	**<DF></DF>**
si plusieurs définitions...	**<DF1></DF1> <DF2></DF2>...**
exemples	**<EX></EX>**
si plusieurs exemples	**<EX1></EX1> <EX2></EX2>...**
catég. grammaticale disjointe	**<CGD></CGD>**
catég. grammaticale jointe	**<CGJ></CGJ>**
figuré	**<FIG></FIG>**
par extension	**<PEX></PEX>**
commentaire encyclopédique	**<CE></CE>**
sous-adresse	**<SA></SA>**
illustration	**<IL></IL>**
métalangage	**<ML></ML>**
adresse monoforme	**<AM></AM>**

Comme le prouve ce tableau, certains choix d'analyse ne sont pas pertinents, surtout en vue d'un balisage XML, nous ne formulerons que deux critiques :

- sur la forme, les balises sont beaucoup trop nombreuses (et ne sont présentées ici qu'une partie) et, de ce fait, ingérables ;
- la rubrique domaine qui suppose de créer une balise par nom de domaine de spécialité « à recenser au fur et à mesure du nettoyage » : procéder de telle sorte n'a aucun sens, une balise de spécialité de domaine suffit et évite une multitude de balises inutiles et fastidieuses à relever.

Pour exemple, voici les entrées « affaiblissement » et « affainéantir » prébalisées :

Exemples de définitions balisées	Signification des balises
<ENT> <AM>AFFAIBLISSEMENT</AM> <PR> (a-fè-bli-se-man) </PR>	ENT entrée AM adresse monoforme PR prononciation

<CG> n. m. </CG> <DF> Diminution de force, d'activité, <EXT> au pr. et au fig. </EXT> </DF> </ENT>	CG catégorie grammaticale DF définition EXT extension d'usage dans la définition CADR complément d'adresse
<ENT> <AM>AFFAINÉANTIR </AM> <CADR> (S') </CADR> <PR> (sa-fé) </PR> <CG> v. pr. </CG> <DF> Devenir mou, lâche. </DF> </ENT>	

Marine Lesprit est chargée du lancement de cette phase : en deux mois, elle balise une cinquantaine de page ; ces pages prébalisées seront en vue du balisage XML. Cette analyse lexicographique sur un échantillon permet de mieux connaître le dictionnaire et d'en modéliser une structure très fine.

4. Travaux en cours

4.1 Automatisation du balisage

Actuellement, nous travaillons parallèlement au nettoyage des fichiers et à l'automatisation du balisage du dictionnaire. De nombreux problèmes se posent pour cette automatisation, et nous les présentons ici, ainsi que les solutions envisagées pour les résoudre.

L'automatisation du balisage est nécessaire : manuellement, il est impossible de baliser plus de dix entrées du dictionnaire par jour, et le dictionnaire en comporte environ quarante mille. Par ailleurs, la réalisation manuelle d'un tel travail nous expose à un risque d'erreur important. L'automatisation se heurte à deux types de problèmes que nous développons maintenant : le premier problème est l'automatisation du balisage des éléments constituant les définitions qui permettra de faire des requêtes dans le dictionnaire, le second est l'automatisation de l'insertion de liens hypertextes entre les définitions.

4.1.1 Balisage du contenu des définitions

Le premier problème auquel nous devons faire face est le problème du balisage des définitions au niveau métalexicographique. En effet, pour permettre une requête, il est nécessaire de baliser le texte pour indiquer à la machine le type d'objet dans lequel elle doit rechercher l'information (définition, exemple, étymologie, etc.). Pour baliser le texte automatiquement à ce niveau, il faudrait que la machine puisse reconnaître, sur la base d'indices fiables, les éléments composant la définition. Ceci semble impossible, en raison des éléments suivants :

Les problèmes de non homogénéité de la rédaction : On trouve par exemple les différences suivantes : il existe deux entrées pour le mot *animal*, une pour l'adjectif, une pour le nom, alors qu'il n'y a qu'une seule entrée pour les deux catégories grammaticales du mot *mammifère*. Nous trouvons deux entrées distinctes pour *âne* et *ânesse*, alors qu'il n'y a qu'une seule entrée pour *chat* et *chatte*. Nous ne pouvons alors pas faire en sorte de créer un programme qui considèrerait qu'il n'y a qu'une seule indication de catégorie grammaticale par entrée, ou encore un programme qui n'insérerait qu'une seule balise pour le genre des noms.

Le problème des marques typographiques identiques pour des informations différentes : On pourrait alors imaginer de se baser sur la typographie pour repérer certaines informations (ceci nécessiterait alors de conserver la mise en forme après l'OCR, ce qui n'est pas évident, mais possible). Cependant, nous observons que :

- Le gras est utilisé pour l'entrée et les proverbes
- L'italique est utilisée pour l'étymologie et les expressions figées
- Les parenthèses sont utilisées pour l'étymologie et la prononciation

On ne peut donc pas espérer récupérer les informations sur la typographie pour permettre la reconnaissance de certains éléments de contenu.

L'absence d'indications formelles pour le passage d'une information à une autre

L'article *mammifère*, par exemple, comporte une définition pour l'adjectif et une définition pour le nom. Les deux définitions sont écrites dans un seul et même paragraphe. Le saut de ligne qui aurait pu aider à délimiter les deux définitions n'était pas présent, on n'a donc encore une fois pas d'indication formelle nous permettant de délimiter des éléments de contenu.

4.1.2 Insertion des liens hypertextes

Il est nécessaire de réaliser le balisage automatiquement, autant pour les requêtes que pour ajouter des liens hypertextes entre les définitions qui permettraient une navigation plus complète.

Par exemple, nous souhaitons qu'en accédant à la définition de *ânesse* (*femelle de l'âne*), l'utilisateur puisse ensuite directement lire la définition de *âne*.

Nous pensons pour l'instant à ne proposer des liens que sur les termes représentant les classificateurs permettant la définition, mais cela pose déjà un certain nombre de problèmes.

Nous souhaitons pouvoir insérer automatiquement les liens dans le texte, pour des raisons de temps, ce qui nous amènera à insérer des balises que nous appellerons source du lien – à l'intérieur des définitions, et des balises cibles – sur la vedette sur laquelle le lien doit pointer. Les problèmes que nous allons rencontrer seront les suivants :

4.1.3 La reconnaissance automatique des classificateurs (balises sources)

Il est impossible pour reconnaître le classificateur, d'utiliser la forme de la définition. Très souvent, il est le premier mot de la définition, mais ce n'est pas toujours le cas (cf. la définition de *ânesse* citée précédemment).

Par ailleurs, nous ne pouvons pas envisager d'utiliser la forme des mots contenus dans la définition. En effet, la forme du classificateur peut varier : il arrive que les formes soient fléchies dans les définitions, ce qui ne sera bien entendu pas le cas dans les vedettes. Ainsi, la définition du mot *mammifère* contient la forme *animaux* et non animal, ce qui rend impossible la réalisation d'un programme informatique basé sur la reconnaissance des formes.

4.1.4 Le problème de l'identification de la cible (vedette sur laquelle le lien doit pointer)

Enfin, reconnaître la vedette cible du lien n'est pas directement possible. Ici se pose le problème des homonymes. Pour la définition de *mammifère*, par exemple, nous souhaitons pointer vers le nom *animal*, qui est le classificateur utilisé dans la définition. Etant donné qu'il existe une entrée différente pour le nom et pour l'adjectif *animal*, un programme basé sur la reconnaissance de lemmes ne sera pas suffisant.

Les problèmes que nous allons rencontrer vont nécessiter pour automatiser le balisage, et en particulier l'insertion de liens hypertextes, une analyse morphosyntaxique du texte avec reconnaissance des lemmes, de manière à pouvoir trouver dans le balisage une information sur les catégories grammaticales et sur la forme de l'entrée quand elle n'est pas fléchie. Ces éléments vont nous permettre d'effectuer une autre forme de prébalisage (on peut parler de couche préliminaire de balisage), au niveau morphosyntaxique, de façon à pouvoir appliquer ensuite des programmes qui baliseront dans un deuxième temps le texte au niveau lexicographique.

5. Conclusion

Bibliographie

AUGE Claude (sous la dir. de) (1906), *Petit Larousse illustré*, Paris : éditions Larousse, 1672 p.

BENAMARA Farah, SAINT-DIZIER Patrick (2003), Dynamic Generation of Cooperative Natural Language Responses . Dans: EWNLG03-ACL , Budapest. ACL , avril 2003.

BONHOMME Patrice (2000), Codage et normalisation de ressources textuelles, *in Ingénierie des Langues*, sous la direction de J-M Pierrel, Hermès, Paris.

GROSS Maurice (1975), *Méthodes en syntaxe* : Paris, Hermann.

ISO (2003)*, Lexical Markup Framework, proposition ISO TC37 / SC4* , accessible à la page : http://pauillac.inria.fr/atoll/RNIL/TC37SC4-docs/N089.pdf

MUELLER M., *A very gentle introduction to TEI*, document Internet accessible à : http://www.tei-c.org/Sample_Manuals/mueller-main.

PRUVOST Jean (2000), *Dictionnaires et nouvelles technologies*, Paris : PUF, 177 p.

SPERBERG-MCQUEEN, CM et BURNARD L (eds), 2004 Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition : http://www.tei-c.org/P4X/index.html

Remerciements

Les auteurs remercient chaleureusement Jean Pruvost, Nicole Cholewka, Anne-Marie Hetzel et Virginie Hababou, pour leur soutien, leurs conseils et pour l'organisation de cette belle journée des dictionnaires.